Studijní opora

BIOSTATISTIKA

Viktor Hynčica, prom.mat.

Rozsah: 10 výukových hodin, 15 hodin konzultací

Úvod

Statistika je v dnešní době významným pomocníkem prakticky ve všech vědních oborech a její použití nemůže chybět ani v oborech biomedicínckých. Dnešní zdravotnictví již není zaměřeno jen na léčení jedince, ale snaží se „léčit“ a ovlivnit zdraví celé populace . V tomto pojetí se musí potýkat s řadou jevů , které mají hromadnou povahu a statistika je právě nástroj jak takovéto jevy hodnotit. Statistika jako pojem již existuje od starověku, kdy byla podobně jako i dnes soustředěna na sociální a ekonomické aspekty vývoje státu. Na začátku 20. století se však od tohoto klasického pojetí odklání díky tomu, že se opírá o pravděpodobnost a stává se samostatnou disciplinou – matematickou statistikou, kterou v oblasti použití ve zdravotnictví či biologie nazýváme biostatistikou.

Cílem předmětu je seznámit studenty se základními pojmy a principy použití biostatistiky v praxi. Dále nebudeme činit rozdíl mezi statistikou a biostatistikou a budeme užívat pouze termín statistika.
Tento text nenahrazuje skripta nebo odbornou publikaci o statistice , pouze stručně seznamuje s okruhy témat ve statistice.

Obsah

1. Základní statistické pojmy

Statistika je věda , jejímž předmětem zájmu jsou hromadná pozorování nebo výsledky opakovaných pokusů (stručně nazýváme jevy) , jejich sběr, analýza a využití pro rozhodování a předpovědi. Statistiku tak lze charakterizovat jako vědu zabývající se studiem hromadných náhodných jevů. Statistika umožňuje na základě výsledků pozorování formulovat obecné závěry oproštěné od subjektivních pohledů, přičemž spolehlivost či nespolehlivost se vyjadřuje pomocí pravděpodobnosti.
Statistické metody dělíme do dvou skupin , jednak na popisnou, neboli deskriptivní statistiku a induktivní statistiku. Popisná statistika se zabývá uspořádáním souboru pozorování, jejich popisem a účelnou sumarizací. Výsledky vyjadřujeme pomocí tabulek, grafů a počítají se základní statistické charakteristiky. Induktivní statistika poskytuje metody , které umožňují z empirických poznatků formulovat obecné závěry pro celou populaci, pro kterou používá induktivní statistika termín základní soubor.
Základní soubor může mít konečný nebo nekonečný rozsah a jeho vlastnosti vyšetřujeme prostřednictvím výběrového souboru, což je část základního souboru, přičemž každý prvek ze základního souboru měl stejnou šanci se dostat do výběrového souboru. Pak takový výběrový soubor nazýváme náhodným, nebo reprezentativním výběrem.
Vlastnosti sledované na prvcích základního souboru nazýváme statistickými znaky. Vlastnosti, které vymezují prvek v souboru jsou znaky určujícími, vlastnosti které jsou předmětem sledování jsou znaky zkoumané.
Podle povahy se dělí znaky na kvalitativní a kvantitativní:

Kvalitativní znaky – obměny u tohoto znaku vyjádřeny pojmem, slovním vyjádřením (pohlaví, vzdělání..)

- alternativní (ano, ne)
- ordinální (lze uspořádat - vzdělání)
- nominální (nelze uspořádat – pohlaví)

Kvantitativní znaky – obměny vyjádřeny polohou na číselné ose (věk, výška…)

- spojité (mohou nabývat jakékoli reálné hodnoty v určitém intervalu – délka, teplota)
- diskrétní (mohou nabývat pouze celá čísla- počty)
-

 Klíčová slova: Deskriptivní, induktivní statistika, znak kvalitativní, znak kvantitativní

 Jakou metodou lze získat náhodný výběr ?

2. Statistické šetření

Statistické šetření představuje základní metodu použití statistiky v praxi a lze je zhruba rozdělit do čtyř etap:

1. plán statistického šetření
2. sběr dat
3. popis souboru
4. rozbor a závěry

Plán statistického šetření představuje důležitou etapu , jejímž cílem je získat data, která lze zpracovat statistickými metodami. Předpokladem pro tuto etapu je jasná formulace problému, stanovení hypotéz a cíle výzkumu. Úkoly při plánování statistického šetření spočívají v zajištění homogenity pozorování a strukturní homogenity. Homogenita pozorování znamená, že se měřené znaky na statistických jednotkách zachycují jednotným způsobem. Strukturní homogenita znamená , že se eliminuje vliv rušivých veličin, známých či neznámých.
Podle rozsahu výběru a cíle statistického šetření se rozhodne pro určitou formu záznamu dat a připraví elektronický dotazník pro vkládání dat ve vhodném statistickém programu pro vytvoření databáze získaných dat. Velice důležité je zajistit programovou kontrolu dat v průběhu zadávání.
Ve vybraném statistickém programu se použijí deskriptivní statistické metody pro prvotní zpracování a případné vyčištění dat od chybných údajů.
Pro vyčištěná data se pak použijí vybrané statistické testy a metody pro vlastní analýzu a hodnocení cílů a hypotéz statistického šetření.

 Klíčová slova: Homogenita pozorování, strukturní homogenost

 Co znamená slepý pokus?

4. Deskriptivní metody statistiky

Pro základní popisné – deskriptivní zpracování dat se používají tři základní metody:

- statistické třídění
- grafické znázornění
- výpočet statistických ukazatelů

Statistické třídění znamená rozdělení souboru dat do skupin neboli tříd podle určených třídících znaků, což umožní poznat v hrubých rysech strukturu a rozložení znaků. Pokud třídíme podle jednoho znaku mluvíme o jednostupňovém třídění, pokud třídíme podle dvou nebo více mluvíme o kombinačním třídění. Znaky sloužící za podklad třídění musí vyjadřovat podstatu zkoumaného jevu a musí být voleny podle cíle výzkumu.
Třídění pro kvalitativní znaky nepotřebuje komentáře, pro kvantitativní znaky je třeba předem určit tzv. třídní intervaly , které jednoznačně pokrývají rozsah výsledků zkoumaného znaku. Následující příklad uvádí třídění pro vitální kapacitu plic u souboru 90ti mužů ve věku 40-50 let. Pro možnost srovnání s jiným souborem jsou v tabulce třídění použity i relativní hodnoty v procentech.

třídní interval	střed intervalu x_i	četnost		kumulativní četnost	
		abs.	relat. v %	abs.	relat. v %
2,75 - 3,24	3,0	1	1,11	1	1,11
3,25 - 3,74	3,5	11	12,22	12	13,33
3,75 - 4,24	4,0	12	13,33	24	26,67
4,25 - 4,74	4,5	22	24,44	46	51,11
4,75 - 5,24	5,0	16	17,78	62	68,89
5,25 - 5,74	5,5	13	14,44	75	83,33
5,75 - 6,24	6,0	10	11,11	85	94,44
6,25 - 6,74	6,5	2	2,22	87	96,67
6,75 - 7,24	7,0	1	1,11	88	97,78
7,25 - 7,74	7,5	2	2,22	90	100,00
Součet		90	99,98		

Na statistické třídění lze snadno navázat i grafické znázornění rozložení hodnot zkoumaných znaků. Grafické znázornění poskytuje rychlou a názornou informaci o zjištěných hodnotách , dovoluje velice názorně srovnat vývoj několika měřených znaků. Pro prezentaci kvantitativních údajů se používá sloupcových grafů, histogramů a polygonů četností. K prezentování struktury sledovaných kvalitativních znaků se nejčastěji používá graf výsečový (kruhový, „koláčový). Při studiu závislostí se používá bodový graf nebo spojnicový graf, ze kterého lze vyčíst typ závislosti, její směr a sílu.
Pro znázornění geografického rozložení četnosti nějakého jevu (např.nemocnosti) pak používáme kartogramy.
Zásadní použití v deskriptivní statistice mají statistické ukazatele, o kterých pojednává následující kapitola.


Klíčová slova: Jednostupňové třídění, kombinační třídění, sloupcový graf, výsečový graf, spojnicový graf, kartogram


Podle jakých znaků je vhodné třídit výskyt nějaké nemoci, abychom se o té nemoci něco dozvěděli?

5. Statistické ukazatelé

Statistické ukazatele poskytují souhrnnou informaci o sledovaných údajích, charakterizují frekvenci sledovaných jevů, nakupení hodnot měřených znaků v určitých místech, jejich kolísání, variabilitu apod. Účelně nahrazují celou množinu pozorování jedním číslem, které charakterizuje určitou vlastnost sledovaného znaku.
Statistických ukazatelů existuje celá řada a jejich použití závisí na tom, zda se jedná o kvalitativní či kvantitativní znak.
Relativní ukazatelé se používají při práci s kvalitativními znaky :

- Strukturální, extenzitní
- Frekvenční, intenzitní
- Indexy

Strukturální ukazatelé charakterizují rozčlenění souboru podle určitého hlediska (pohlaví, věk, vzdělání ..) a jde o procenta či promile:

$$\frac{\text{abs.četnost\_jevu}}{\text{rozsah\_souboru}} \quad , \quad N = 100 \text{ procento}, \quad N = 1000 \text{ promile}$$

Frekvenční ukazatelé vyjadřují častost výskytu nějakého jevu, např.frekvenci výskytu nemoci:

$$\frac{\text{poč.nemocných}}{\text{celk.poč.obyvatel}} \cdot 100.000$$

Indexy se používají k posouzení vývoje v čase nějakého ukazatele. Jde o indexy s pevným základem, které charakterizují celkový trend, nebo indexy s pohyblivým základem, které charakterizují dynamiku vývoje.

Při zpracování kvantitativních znaků používáme nejčastěji ukazatelů polohy a variability pro popis chování znaku.
Ukazatelé polohy, nebo též střední hodnoty, informují o poloze rozložení četností na reálné ose a tedy i o velikosti sledované veličiny. K nejběžnějším používaným ukazatelům polohy patří aritmetický průměr, medián a modus.

Aritmetický průměr:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

kde X_i značí jednotlivá pozorování

Medián je hodnota prostředního prvku souboru, který je uspořádán podle velikosti

Modus je nejčastěji se vyskytující obměna znaku v daném souboru, tedy hodnota s největší četností.

Ukazatele polohy poskytují velmi důležitou, ne však úplnou a dostačující informaci o souboru kvantitativních údajů. Kromě koncentrace údajů kolem středních hodnot je třeba charakterizovat i jejich variabilitu, která se projevuje kolísáním údajů kolem ukazatelů polohy v určitých vzdálenostech

Základní ukazatel variability je rozptyl, který měří variabilitu pomocí odchylek jednotlivých údajů od průměru. – je to průměr čtverců těchto odchylek:

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

Rozptyl je z teoretického hlediska základním ukazatelem variability, z praktického hlediska je těžko interpretovatelný. Proto se zavedl jako další ukazatel variability ukazatel směrodatná odchylka:

$$S.D. = s = \sqrt{s^2}$$

Výhodou směrodatné odchylky je její věcný význam, neboť nám dává představu o tom jak jsou údaje rozptýleny v rámci celého souboru. V případě tzv. normálního rozdělení znaku se v rozmezí dvojnásobné vzdálenosti směrodatné odchylky od průměru nachází přibližně 95% všech údajů.

Pokud potřebujeme srovnat variabilitu ve dvou, nebo několika souborech, jejichž průměry se značně liší, nebo variabilitu znaků uváděných v různých měrných jednotkách (výška, hmotnost) používáme ukazatel nazývaný variační koeficient:

$$x \qquad \frac{S.D.}{\bar{x}}$$

Kromě těchto základních ukazatelů pro kvantitativní znaky existují i další, které charakterizují „tvar“ rozložení znaků . Jde např. o šikmost a špičatost.

Klíčová slova: Extenzitní, intenzitní ukazatele, indexy, střední hodnoty, průměr, median, modus, rozptyl, směrodatná odchylka, variační koeficient.

Typickým frekvečním ukazatelem je nemocnost na 100.000 obyvatel. Někdy ho při porovnávání zkreslí rozdílná věková struktura v porovnávaných souborech. Jak lze eliminovat vliv věku , co je to standardizace nemocnosti?

6. Základy pravděpodobnosti.

Teorie pravděpodobnosti podstatně ovlivnila vývoj statistiky a způsobila, že dnes je matematická statistika významným pomocníkem v mnoha vědních oborech . Důvodem toho je , že dovede popsat a pracovat s náhodným projevem jevů v přírodě. Dnes se pravděpodobnost opírá o matematické discipliny jako o teorii množin a teorii míry a není už tou klasickou pravděpodobností jako v počátcích, kdy se začala vyvíjet na pozadí hazardních her.

Pravděpodobnost je spjata s pojmy náhodný pokus a náhodný jev. Pravděpodobnost je kvantitativní charakteristika, která je mírou častosti výskytu náhodného jevu.
Klasická pravděpodobnost předpokládá, že náhodný pokus má konečný počet výsledků, tak zvaných elementárních jevů. Náhodný jev lze chápat jako libovolnou množinu elementárních jevů a pravděpodobnost je pak definována jako podíl počtu elementárních jevů příznivých náhodnému jevu ku celkovému počtu všech elementárních jevů.
$P(A) = m / n$, m= počet elementárních jevů příznivých jevu A, n= počet všech elementárních jevů

Základní vlastnosti pravděpodobnosti:

$0 \leq P(A) \leq 1$ $P(A) = 0$ …jev nemožný, $P(A) = 1$ … jev jistý

Věta o sčítání pravděpodobností:

$P(A \cup B) \leq P(A) + P(B)$ jedná se o pravděpodobnost spojení dvou jevů , platí však pouze tehdy, jestliže jsou jevy A a B disjunktní, tedy nemohou nastat současně. Pokud tomu tak není , je nutné odečíst pravděpodobnost současného výskytu těchto jevů:

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Věta o násobení pravděpodobností:

$P(A \cap B) = P(A) \cdot P(B)$ jedná se o pravděpodobnost současného výskytu jevů A a B, $P(B)$

platí však pouze v případě , že jevy A a B jsou nezávislé , jinak je nutné zavést podmíněnou pravděpodobnost P(B/A):

P(A □ □B) □ □P(A) □ □P(B / A) , P(B/A) značí pravděpodobnost jevu B pokud nastal jev A Další vlastnosti a teorie pravděpodobnosti přesahuje rámec tohoto dokumentu.

 Náhodný jev, náhodný pokus, pravděpodobnost, podmíněná pravděpodobnost, nezávislost jevů

 Jaká je pravděpodobnost, že se narodí po sobě dva kluci ?

7. Teoretické rozložení

V rámci deskriptivní statistiky lze popsat výběrová rozložení kvantitativních znaků velikostí výběru, průměrem, směrodatnou odchylkou a například histogramem četností, pro kvalitativní znaky pak rozsahem výběru a relativními ukazateli-četnostmi. Tyto ukazatele , které počítáme z výběru , jsou s daným výběrem svázány a nazýváme je výběrovými charakteristikami. Matematická statistika zavádí pro rozložení hodnot znaků v celém základním souboru (populaci) modely, které jsou popsány teoretickým rozložením založeném na frekvenční křivce(hustotě pravděpodobnosti) v případě kvantitativních znaků s parametry které odpovídají průměru a rozptylu výběrového souboru. V případě kvalitativních znaků je model popsán pravděpodobnostní funkcí, což je výčet pravděpodobností s jakými se vyskytuje každá možná hodnota znaku.
Mezi výběrovými charakteristikami a parametry základního souboru je zásadní rozdíl. Parametry jsou pro daný základní soubor neměnné konstanty, zatímco výběrové charakteristiky jsou jednoznačně určeny vždy pro daný konkrétní výběr. Přestože v přírodním světě panuje velká různorodost a mnohotvárnost, ukazuje se , že při popisu reálných náhodných jevů lze vystačit s poměrně malým počtem teoretických rozložení. Tato teoretická , modelová rozložení jsou po matematické stránce dobře propracovaná, jsou stanoveny rovnice jejich frekvenčních křivek , či pravděpodobnostních funkcí.
Pro modelování kvantitativních spojitých znaků se nejčastěji používá Normální- Gaussovo rozložení, z dalších pak Studentovo, Snedecorovo ,či Chí ^{2} . Pro diskrétní znaky je pak nejfrekventovanější Binomické a Poissonovo rozložení.

Normální – Gaussovo rozložení.

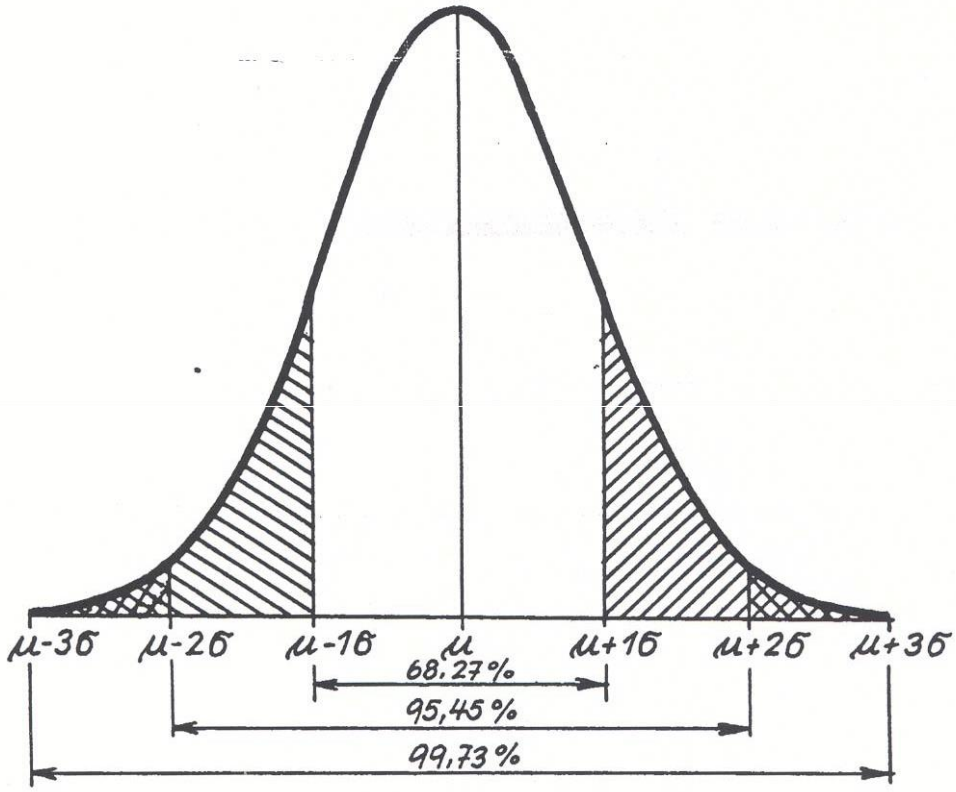
Toto rozložení je nejdůležitějším teoretickým rozložení pro hodnocení spojitých kvantitativních znaků. Je možné ukázat, že znak má normální rozložení , vzniká-li současným působením velkého počtu nepatrných, nezávislých příčin nahodilého charakteru. Frekvenční křivka je definována rovnicí:

□□(x□□□)^2 □
□□_____□
1 □ □
_____ e □ □
□ √ 2□

, kde μ a σ jsou parametry normálního rozložení - průměr

a směrodatná odchylka, které jednoznačně normální rozložení definují, π a **e** jsou konstanty.

Následující graf ukazuje průběh frekvenční křivky normálního rozložení spolu s praktickým významem směrodatné odchylky.



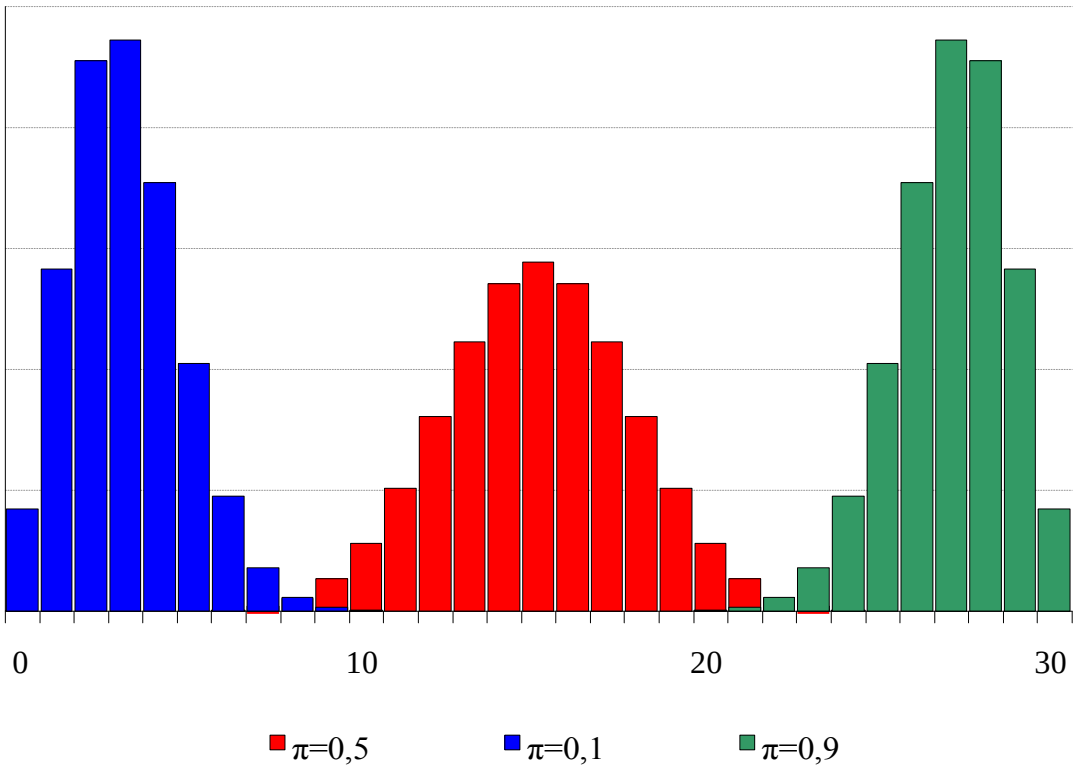
Binomické rozložení

Binomické rozložení je rozložení četností výskytu jevu , který může nabývat pouze dvou hodnot. Typickým příkladem může být pokus s úhynem myší po určité nákaze. Pokud je pravděpodobnost úhynu po nákaze π a v pokusu máme například 30 myší, které byly nakaženy, binomické rozložení definuje pravděpodobnost , že z 30ti myší zahyne **x.** Tato pravděpodobnost je definována :

$$P(x) \quad \square\square\square \quad \square\square \; x \quad 30\square\square x \quad (1\square\square\square)$$

$$\square \quad x \; \square$$

a následující graf zobrazuje všechny možnosti pro 3 různé hodnoty pravděpodobnosti π :



Obecná definice binomického rozložení je:

n

$$P(x) \quad \quad x^{x} (1 \quad \quad)^{n \quad x}$$

$$\quad \quad \quad X$$

$$\quad \quad$$

$$\quad \quad$$

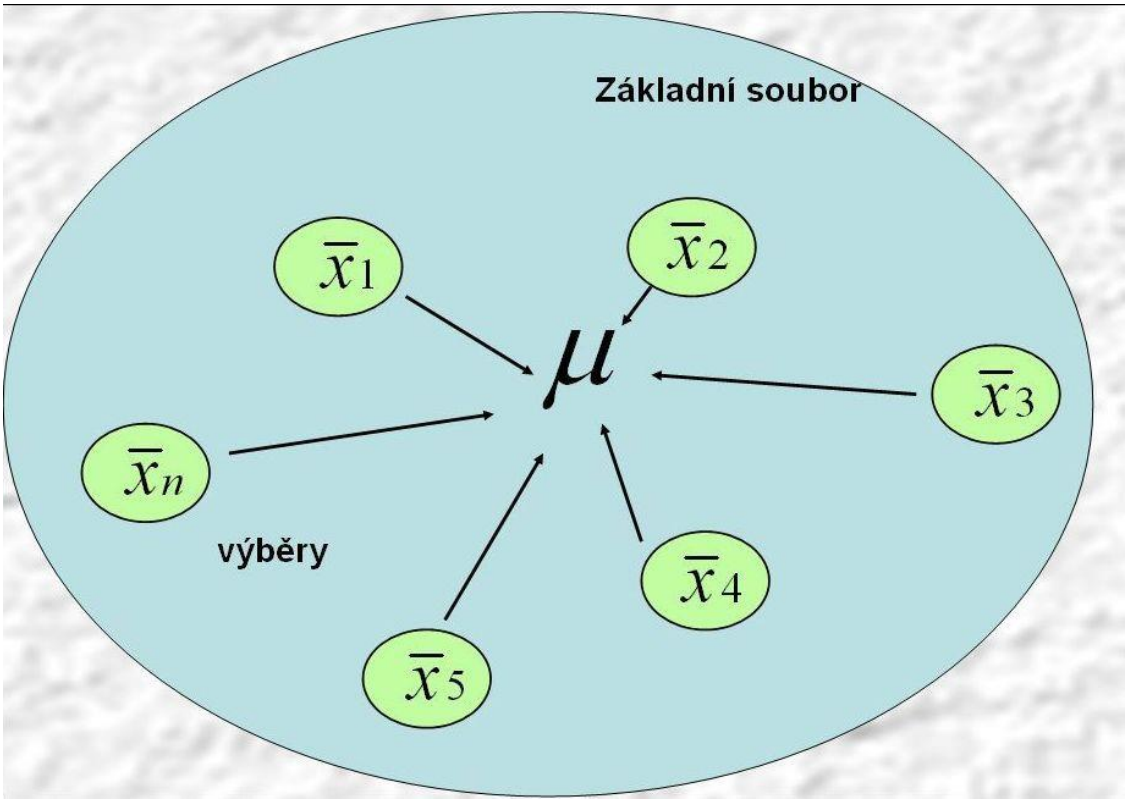
Klíčová slova: Výběrové charakteristiky, parametry, frekvenční křivka , Gaussovo normální rozložení, binomické rozložení.

V čem spočívá rozdíl mezi výběrovými charakteristikami a parametry?

8. Statistické odhady.

Cílem teorie statistických odhadů je odhadnout neznámé parametry základního souboru pomocí výběrových charakteristik. Pokud odhadujeme neznámý parametr jedním číslem jedná se o **bodové odhady,** např. odhad průměru základního souboru výběrovým průměrem. Většinou se však užívají **intervalové odhady** , které se vytváří tak, že kolem bodového odhadu utvoříme interval, který s předem zvolenou vysokou pravděpodobností (95%, 99%) pokryje odhadovaný parametr. Jedná se intervaly spolehlivosti 95-procentní nebo 99-procentní a nazývané též konfidenčními intervaly.

Jako příklad uvedeme konstrukci intervalu spolehlivosti pro průměr základního souboru, který ilustruje následující obrázek:



Konstrukce vychází z vlastností výběrových průměrů. Pokud má náhodná kvantitativní veličina normální rozložení s parametry **μ** a **σ** , pak výběrové průměry z náhodných výběrů rozsahu alespoň **n** >=30 mají také normální rozložení s parametry **μ** a **σ**/√ **n** . Směrodatná odchylka výběrových průměrů je tedy √n -krát menší nežli je směrodatná odchylka náhodné veličiny a nazývá se **střední chybou průměru**..
Z vlastností normálního rozložení pak lze snadno odvodit intervaly spolehlivosti:

95ti-procentní interval spolehlivosti:

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n-1}}$$

99ti-procentní interval spolehlivosti:

$$\bar{x} \pm 2.58 \frac{s}{\sqrt{n-1}}$$

Analogicky je možné odvodit i intervalový odhad pro neznámou pravděpodobnost nějakého jevu.

Klíčová slova: Bodový odhad, intervalový odhad, konfidenční interval, střední chyba průměru.

Pro konstrukci konfindenčního intervalu pro průměr jsme museli nahradit střední chybu průměru $\sigma/\sqrt{n}$ bodovým odhadem (proto se ve vzorci pod odmocninou vyskytuje n-1).

9. Testování hypotéz.

Teorie testování hypotéz je zásadním nástrojem použití matematické statistiky ve výzkumu. Použitím teorie testování hypotéz můžeme prověřovat nejrůznější předpoklady a domněnky, které vyplývají ze stanovených cílů výzkumu. Například, zda se 2 soubory liší v průměru nějakého znaku, zda výskyt nějaké nemoci závisí na nemoci, zda lék A je lepší než B atd. Statistická hypotéza je vlastně výrok o statistickém souboru , o zkoumaném pravděpodobnostním rozložení. Platnost statistických hypotéz se prověřuje na základě pozorovaných dat , pomocí různých statistických testů zvaných testy významnosti. Prověřovaný předpoklad označujeme jako hypotézu H_0 , tedy jako nulovou hypotézu a statistický test významnosti je kriterium , které na základě pozorovaných dat dovoluje jednoznačně rozhodnout zda se nulová hypotéza H_0 zamítne nebo nezamítne. Hypotézu ke které se přikloníme po zamítnutí H_0 nazýváme alternativní hypotézou A.
Mohou nastat tyto čtyři případy:

- Hypotéza H_0 platí a nebyla testem zamítnuta
- Hypotéza H_0 platí a byla testem zamítnuta – chyba 1. druhu
- Hypotéza H_0 neplatí a nebyla testem zamítnuta – chyba 2. druhu
- Hypotéza neplatí a byla testem zamítnuta

Pravděpodobnost chyby 1. druhu se nazývá hladinou významnosti, značíme ji α a u jednotlivých testů se zjišťuje zda je pod zvolenou hranicí, obvykle 0,05 nebo 0,01. Pravděpodobnost chyby 2. druhu rozhoduje o síle testu.
Prakticky se test provádí tak, že zvolíme vhodnou testovací charakteristiku, což je určitá funkce F jednotlivých pozorování, která má za platnosti testované hypotézy známé rozložení a dává tak možnost rozhodovat mezi testovanou a alternativní hypotézou. Vypočtenou hodnotu funkce F pak porovnáme s kritickou hodnotou testovací statistiky (pro α =0.05 nebo 0,01). Pokud testovací statistika F je větší než odpovídající kritická hodnota , je pravdivost hypotézy H_0 málo pravděpodobná a proto ji zamítáme. V opačném případě je možné vysvětlit pozorované rozdíly prostřednictvím náhody a hypotézu tedy nezamítáme. Neznamená to však její přijetí, statistickými testy můžeme testované hypotézy pouze vyvracet , nikoli dokazovat jejich platnost.
Statistické metody pro testování hypotéz lze rozdělit do dvou hlavních skupin:

- Metody parametrické
- Metody neparametrické-pořadové

Metody parametrické jsou založeny na známém rozložení náhodných veličin, většinou normálním, a používají se pro testování hypotéz výběrových parametrů. Metody neparametrické nepředpokládají žádnou znalost rozložení veličin a využívají například pouze pořadí zjištěných hodnot.

Statistických testů ať parametrických či neparametrických bylo odvozeno mnoho, dále si jako příklad uvedeme test , zda se 2 soubory liší v průměru nějakého znaku.

1.soubor…… $n1, \bar{x}1, s1$
2.soubor……. $n2, \bar{x}2, s2$

Předpoklady: $n1, n2 > 30$, rozložení dat zhruba normální ,stejné rozptyly, nezávislé výběry

Hypotéza H_0 : $\sigma_1 = \sigma_2 = \sigma$ Hypotéza A : $\sigma_1 \neq \sigma_2$ Za platnosti H_0 :

rozdíly $(x_1 - x_2)$ kolísají kolem nuly podle normálního rozložení se směrodatnou odchylkou - střední chyba rozdílu dvou průměrů :

$$s_{\bar{x}1-\bar{x}2} = \sqrt{s_{\bar{x}1}^2 + s_{\bar{x}2}^2}$$

Z vlastností normálního rozložení plyne ,že s pravděpodobností 0,95 leží rozdíly v intervalu:

$$\bar{x}1 - \bar{x}2 \pm 1.96s$$

A můžeme definovat pro srovnání 2 průměrů U-test:

$$u = \frac{\left| \bar{x}_1 - \bar{x}_2 \right|}{s_{\bar{x}1-\bar{x}2}}$$

Interpretace výsledků:

u > 1.96 rozdíl je statisticky významný na hladině $\alpha = 0.05$
u > 2.58 rozdíl je statisticky významný na hladině $\alpha = 0.01$

Tato podoba platí pro výše uvedené předpoklady, pro obecnější podmínky lze nalézt v literatuře úpravu tohoto testu jako **t-test**.

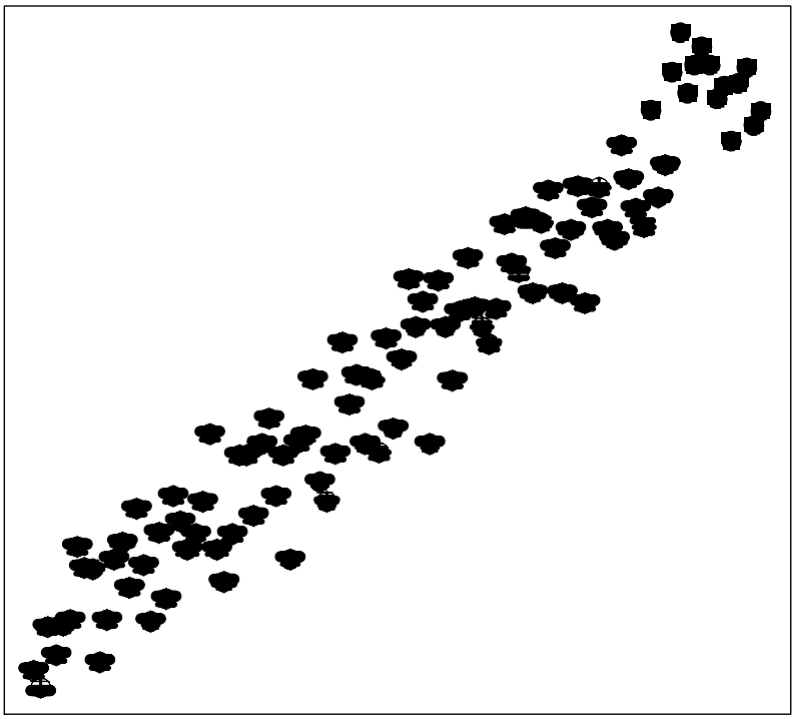
Pro podobné situace lze používat i neparametrické testy – mediánový a Wilcoxonův. Mediánový test hodnotí, zda jsou častěji nad mediánem společného souboru pozorování z 1. souboru než z 2. souboru. Wilcoxonův test porovnává průměrné pořadí prvků z prvního souboru s průměrným pořadím prvků z druhého souboru v souboru , který vznikl spojením obou.

Klíčová slova: Nulová hypotéza, alternativní hypotéza, hladina významnosti, statistická signifikance, kritická hodnota, parametricý test, neparametrický test, t-test, Chi^2 –test.

Kdy nastává chyba prvního nebo druhého druhu?

10. Závislost kvantitativních dat.

Typické pro statistiku je také to, že sledujeme více znaků a hodnotíme jejich závislost. Statistické metody používané k posouzení závislosti jsou jiné pro kvantitativní a jiné pro kvalitativní znaky. V této kapitole se věnujeme závislosti pro kvantitativní znaky. Statistická závislost pro kvantitativní znaky se liší od závislosti , kterou známe z fyziky, kdy jedné hodnotě jedné veličiny odpovídá jediná hodnota druhé veličiny , např. závislost ujeté vzdálenosti na čase u rovnoměrného přímočarého pohybu – $s = v \cdot t$.
U statistické závislosti jedné hodnotě jedné veličiny odpovídá celý soubor hodnot druhé veličiny. U každé závislosti rozlišujeme typ závislosti a těsnost. Obojí můžeme snadno posoudit z bodového grafu:



Typ závislosti určuje čára, která lze proložit body v grafu, vyjadřuje se matematickou funkcí a nazýváme jí regresní funkcí. Rozlišujeme závislosti lineární (proložená čára je přímka), dále např. logaritmické exponenciální, parabolické atd. Přiléhavost a vhodnost vybraného typu křivky lze ověřovat vhodnými statistickými metodami.
Těsnost závislosti, tj. rozptýlenost bodů kolem proložené čáry, lze posoudit z grafu a posoudit statistickými testy. Dále se omezíme jen na lineární závislosti a můžeme

hovořit o lineární regresní analýze. Model lineární regresní analýzy je definován rovnicí:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

, kde y_i a x_i je i-té pozorování závislé, respektive nezávislé proměnné, α a β jsou parametry rovnice přímky a ε_i je odchylka od modelu , neboli reziduum.
Parametry α a β se určí tak, aby tzv.reziduální rozptyl byl minimální:

$$s_e = \frac{1}{N-2} \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2$$

Těsnost lineárního vztahu se posuzuje Pearsonovým koeficientem korelace **r** :

$$r = \frac{\sum_{i=1}^{N} x_i y_i - \bar{x} \sum y_i - \bar{y} \sum x_i}{\sqrt{\left(\sum x_i^2 - \bar{x} \sum x_i\right)\left(\sum_{i=1}^{N} x_i^2 - \bar{x} \sum x_i\right)}}$$

Vlastnosti korelačního koeficientu:

- Nabývá hodnot z intervalu -1 až 1
- Je-li roven nule, jsou znaky nezávislé
- je kladný pro případ přímé závislosti
- je záporný pro případ nepřímé závislosti
- v případě r= 1 nebo r= -1 jde o funkční závislost

Čím více se korelační koeficient blíží v absolutní hodnotě k 1 , je závislost těsnější. Statistickým testem lze hodnotit , zda korelace pro daný případ je významná, tedy že korelační koeficient se významně liší od 0.

Klíčová slova: lineární , nelineární závislost, regresní analýza, korelační koeficient

11. Závislost kvalitativních dat.

Potřeba hodnotit závislost kvalitativních znaků je ve zdravotnictví nebo v lékařském výzkumu velice častá. Příkladem může být potřeba studovat , zda výskyt nějaké nemoci souvisí s výskytem nějakého nepříznivého vlivu, např. kouření. Výběr metod pro hodnocení závislosti kvalitativních znaků závisí na tom, kolika obměn znaky nabývají. Pro znaky, které nabývají více obměn než 2, hodnotíme závislost chi-kvadrát testem χ^2.
Výsledky v tomto šetření se zapisují do kontingenční tabulky, jejíž rozměr odpovídá rozsahu obměn jednotlivých znaků. Příkladem je následující tabulka , která ukazuje vztah typu nádoru na jeho lokalizaci v souboru 152 pacientů, přičemž četnosti v jednotlivých buňkách bez závorek jsou skutečně pozorované četnosti.

Lokalizace nádoru	Typ nádoru			Celkem
	B_1	B_2	B_3	
A_1	28 (24,6)	10 (11,3)	6 (8,1)	44
A_2	22 (16,8)	4 (7,7)	4 (5,5)	30
A_3	35 (43,6)	25 (20,0)	18 (14,4)	78
celkem	85	39	28	152

V tomto případě je hypotézou H0 tvrzení, že typ nádoru nezávisí na jeho lokalizaci a za tohoto předpokladu jsou vypočteny četnosti tzv. očekávané, které jsou v tabulce v závorkách. Ty se odvodí na základě věty o násobení pravděpodobností pro dva nezávislé jevy:

P(A1∩B1) = P(A1) . P(B1) ……..atd.

Odhad pro tuto pravděpodobnost je : **(44 / 152) . (85 / 152)**

Očekávaná četnost pak: **(44/152) . (85/152) . 152 = 24,6**

Chi-kvadrát test je definován vztahem:

$$\chi^2\text{-test nezávislosti} = \sum \frac{(\text{pozorované} - \text{ocekavané})^2}{\text{ocekavané}}$$

V našem případě je $\chi^2 = \sum \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 9{,}25$

Platnost testované hypotézy se prověří srovnáním s kritickou hodnotou - závisí na velikosti tabulky:

(a-1) . (b-1)	kritické hodnoty	
	5%	1%
1	3,84	6,63
2	5,99	9,21
3	7,81	11,3
4	9,49	13,3
5	11,1	15,1
6	12,6	16,8
7	14,1	18,5
8	15,5	20,1
9	16,9	21,7
10	18,3	23,2
11	19,7	24,7
12	21,0	26,2
13	22,4	27,7
14	23,7	29,1
15	25,0	30,6

a = počet řádků tabulky, b = počet sloupců tabulky (bez součtů)

Protože vypočítaná hodnota □2 je menší než 5%-ní kritická hodnota (9,49) testovaná hypotéza se nezamítá.

V lékařském výzkumu často analyzujeme závislost , kterou charakterizuje kontingenční tabulka 2 x 2. Jde o případ , kdy sledujeme, zda závisí výskyt nemoci na expozici rizikovým faktorem, tedy např. na kouření.

Obecný tvar kontingenční tabulky:

Nemoc

Ano Ne

Expozice	Ano	*a*	*b*	a+b
	Ne	*c*	*d*	c+d
:		a+c	b+d	

V této situaci zavádíme míry , které hodnotí riziko výskytu nemoci v závislosti na expozici. Nejznámější míry jsou **Relativní riziko** a **Poměr šancí**.

Relativní riziko se definuje jako poměr Rizika exponovaných ku Riziku neexponovaných., tedy:

$$RR = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = \frac{a(c+d)}{c(a+b)}$$

$RR = R_{exp} / R_{neex}$

Podobně poměr šancí se definuje jako poměr **šance na onemocnění exponovaných** ku **šanci na onemocnění neexponovaných**, tedy:

$$OR = O_{exp} / O_{neexp} \qquad OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

Zkratka OR vychází z anglického pojmenování – odds ratio.

Pokud je RR= 1 = OR nezávisí nemocnost na expozici.

Interpretace např. pro OR je:
Odds ratio = 1 … nezávisí onemocnění na expozici
Odds ratio > 1 … pozitivní asociace – větší šance onemocnět u exponovaných
Odds ratio < 1 … negativní asociace – větší šance onemocnět u skupiny neexponovaných

Klíčová slova: Kontingenční tabulka, očekávaná četnost, χ^2-test nezávislosti, relativní riziko, poměr šancí.

Literatura

J.Hendl: Přehled statistických metod zpracování dat. Portál, Praha 2004.
K. Zvára: Biostatistika. Karolinum, Praha 1998.
Eileen Magnello, Borin Van Loon: STATISTIKA. Portál, Praha 2010.
Marek Luboš a kolektiv: Statistika v příkladech. Professional publishing, Praha 2013